CNES Clearinghouse Prototype Description

Context

Part of CNES R&D activity related to EO data systems

Two objectives :

- understand the problem of attaching semantics to datasets in addition to syntax
- experiment state of the art semantical information representation techniques

Way to reach the objectives :

- develop a prototype and test the resulting concepts on real datasets

Background :

- good experience with syntax attached to datasets through the development of the "EAST" language
- good involvment of CNES in the definition of the CCSDS "Reference Model for an Open Archival Information System"
- good involvment of CNES in the definition of the CCSDS "Data Entity Dictionary Specification Language"

Preliminary steps

Vocabulary clarification

- information vs data
- syntactical analysis vs semantical analysis
- concept vs metadata

Analysis of the work carried out by digital libraries

- SGML
- Text Encoding Initiative

Discovery of the ISO world

- ISO 1087 – *Terminology – Vocabulary*
- ISO 11179 – *Specification and standardization of data elements*
- ISO 2788 – *Guidelines for the establishment and development of monolingual thesauri*

Theoretical study on information representation

- information modelling vs information encoding (UML vs XML)

Analysis of user needs

More than 40 different user needs were identified, e.g. :

- need for many user profiles
  - examples : mapping, agriculture
  - a profile definition results in a set of a priori selection criteria
- possible link of metadata to one or more quicklook datasets and/or documents
- need to track dataset history within metadata attached to them
- need to handle personal information attached to metadata on a global basis
- need to have thesauri and dictionaries available for vocabulary control
- need for various selection criteria, e.g.
  - geographic zone
  - beginning and end time of observation
  - period of observation (or season)
  - data content (e.g. geographic feature or attribute)

Other requirements

Compliance with the CCSDS "OAIS" model

- common services
- ingest
- archival storage
- data management
- administration
- access

Seamless user access to similar clearinghouses

- user queries are forwarded to other clearinghouses
  - user gets lists of all available metadata
  - user gets complete metadata wherever it is located

Compliance with major metadata standards

- FGDC or ISO 19115 : ISO 19115 elected (DIS)
- DocBook or Text Encoding Initiative : "TEI lite" elected

Clearinghouse Definition

General definition

"The clearinghouse is an archive for metadata defined with regard to various standards providing online access to thematic information through a web interface."

Available Services

- **main services**
  - **(meta)data storage**
  - **(meta)data ingest**
  - **(meta)data access**
- **ancillary services**
  - **clearinghouse administration**
  - **(meta)datamanagement**

Roles (1/2)

Administrator

- ingestion / extraction / visualization of
  - metadata and associated datafiles (quicklooks)
  - documents
  - thesauri
- document ingestion / modification/ visualization of
  - “personal identification”

Manager

- profile definition by establishing a list of :
  - applicable “topic categories”
  - applicable keywords from designated thesauri
  - applicable selection criteria (feature type, periode, etc.)
  - metadata elements to be presented to the user as a summary of the complete metadata and
  - the presentation style of this summary

Roles (2/2)

User

- indicates his profile (or a generic profile)
- sees the selection criteria associated with the profile
- chooses all or part of the criteria available to him
  - provides additional information (e.g. coordinates of surrounding geographic box)
- gets a list of matching metadata existing on all connected clearinghouses
  - list is sorted by clearinghouses
  - list shows summary metadata elements as defined by the manager
- refines his selection if needed
- may select a specific metadata and see the complete metadata
  - metadata is displayed as a structured text document
  - associated geographic boxes are displayed on a small map
  - associated datafiles may be saved in a local folder
- may select a specific metadata and save it in a local folder
- may select a document or thesaurus and save it in a local folder

Behind the scene

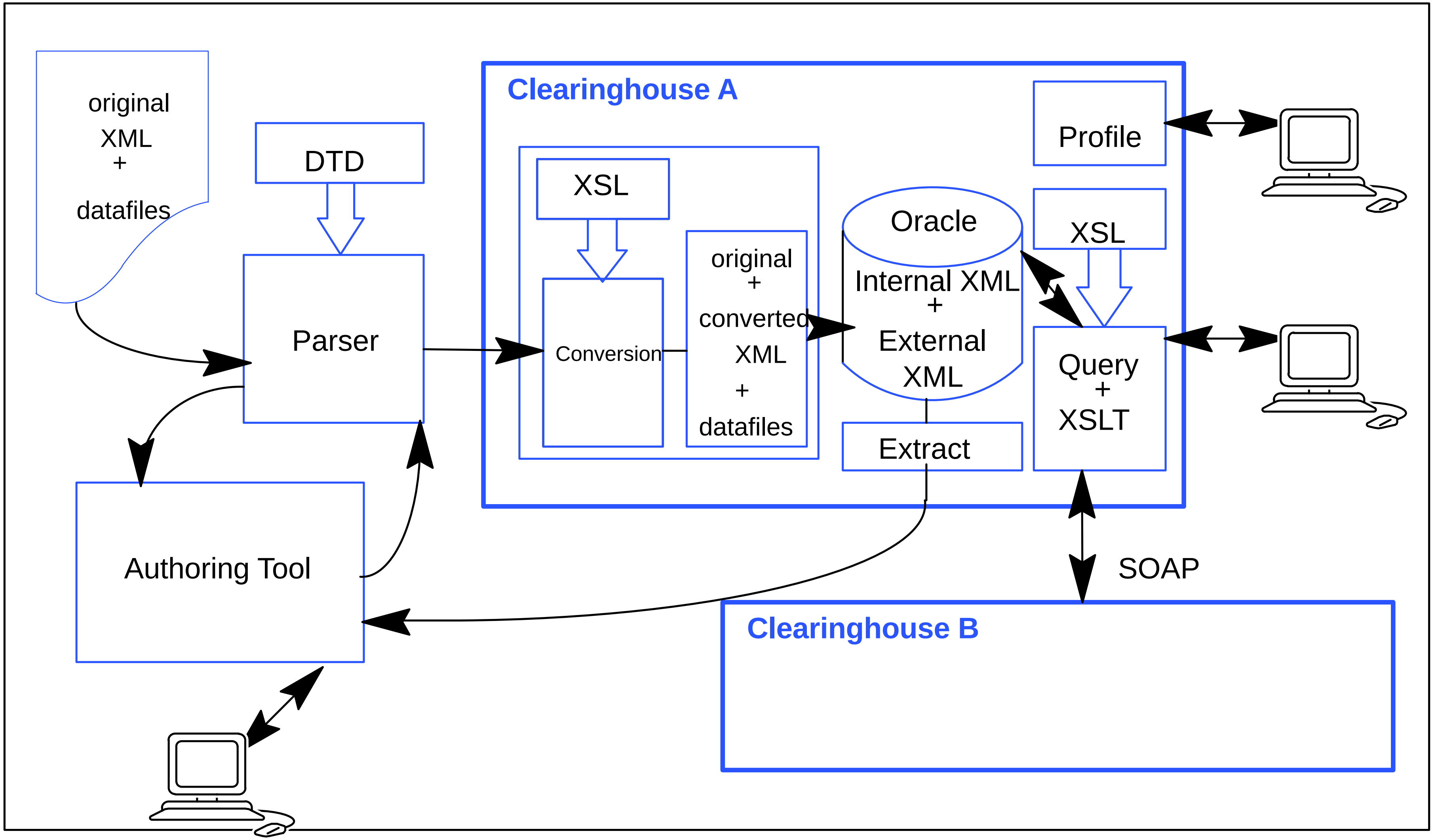
Fundamental assumptions

- clearinghouse is totally “XML” based
  - including the metadata database (ORACLE 9i)
- interoperability is achived via “SOAP”
- heart of the system is a so called “internal DTD”
- “external DTDs” are mapped against this “internal DTD”

Implications

- need to get DTDs from outside
  - Text Encoding Inititaive (lite version)
- need to derive DTDs from UML models
  - ISO 199115 and companions are UML based
- need to build DTDs from scratch
  - thesaurus, etc
- need to get good XML authoring tools

Inside the clearinghouse



Environment

Server

- database : Oracle 9i with XML support
- application : Apache 1.3 with SOAP 2.2 support
- web : Apache Tomcat 4.0.1 with Servlet 2.3 and JSP 1.2 support
- java : jdk 1.3.1, xalan 2.2 (XSLT), xerces 1.4.3 (parser)

Client

- XML authoring tool
  - TurboXML
  - MetalD
  - Quicksilver
- Browser
  - IE5+
  - Netscape 4.7

Next steps

4 months qualification

- installation on CNES Intranet
  - standalone
- qualification
  - with metadata providers (setup of TC211 metadata)
  - with future clearinghouse users
- get feedback

3 additional months qualification

- installation in Ifremer (French sea research institute) premises
- interoperability qualification

Crosswalks to other standards

- DIF
- DIMAP

Expected problems

- ISO 19115 is a complicated standard
- metadata ingest process likely to be difficult